NPL ✓ ADONIS___ MIC___ BioT___Main__NO__Vol NO__ NOS___CKCite__Dup__Int WK

2/23

STIC-ILL

From: Davis, Minh-Tam
Sent: Wednesday, February 22, 2006 5:14 PM
To: STIC-ILL
Subject: Reprint request for 08/785532

1) **The mosaic genome of warm-blooded vertebrates.**
Bernardi G; Olofsson B; Filipski J; Zerial M; Salinas J; Cuny G; Meunier-Rotival M; Rodier F
Science (UNITED STATES) May 24 1985, 228 (4702) p953-8, ISSN 0036-8075 Journal Code: 0404511

2) **enomic analysis of prostate tumors using whole genome and 16q contig array CGH and** genome cryptographer
AUTHOR: Watson Vivienne (Reprint); Kowbel David; Paris Pamela; Andaya Armann; Volik Stanislav; Kamkar Sherwin; James Karen; Sudilovsky Daniel; Schmitt Lars; Shuman Marc; Carroll Peter; Doggett Norman; Rosenburg Carla ; van Dekken Herman; Gray Joe; Albertson Donna; Pinkel Daniel; Collins Colin
AUTHOR ADDRESS: UCSF Cancer Center, San Francisco, CA, USA**USA
JOURNAL: Proceedings of the American Association for Cancer Research Annual Meeting 43 p1067 March, 2002 2002
MEDIUM: print
CONFERENCE/MEETING: 93rd Annual Meeting of the American Association for Cancer Research San Francisco, California, USA April 06-10, 2002; 20020406
ISSN: 0197-016X

3) **Genome sequence and splice site analysis of low-fidelity DNA polymerases H and I involved in replication of damaged DNA.**
Cleaver J E; Collins C; Ellis J; Volik S
UCSF Cancer Center, Box 0808, Room N431, University of California at San Francisco, San Francisco, CA 94143-0808, USA. jcleaver@cc.ucsf.edu
Genomics (United States) Nov 2003, 82 (5) p561-70, ISSN 0888-7543 Journal Code: 8800135

Thank you.
MINH TAM DAVIS
ART UNIT 1642, ROOM 3A24, MB 3C18
272-0830

ACADEMIC
PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

GENOMICS

Genomics 82 (2003) 561–570

www.elsevier.com/locate/ygeno

Genome sequence and splice site analysis of low-fidelity DNA polymerases H and I involved in replication of damaged DNA

J.E. Cleaver,[a,b,*] C. Collins,[a] J. Ellis,[c] and S. Volik[a]

[a] UCSF Cancer Center, Box 0808, Room N431, University of California at San Francisco, San Francisco, CA 94143-0808, USA
[b] Department of Dermatology, University of California at San Francisco, San Francisco, CA 94143-0808, USA
[c] Division of Bioengineering and Physical Science, Office of Research Services, Office of the Director, National Institutes of Health, Building 13, Room 3E-49, 9000 Rockville Pike, Bethesda, MD 20892-5766, USA

Received 30 January 2003; accepted 2 June 2003

Abstract

POLH and *POLI* are paralogs encoding low-fidelity, class Y, DNA polymerases involved in replication of damaged DNA in the human disease xeroderma pigmentosum variant. Analysis of genomic regions for human and mouse homologs, employing the analytic tool Genome Cryptographer, detected low-repetitive or unique regions at exons and other potential control regions, especially within intron I of human *POLH*. The human and mouse homologs are structurally similar, but the paralogs have undergone evolutionary divergence. The information content of splice sites for human *POLH*, the probability that a base would contribute to splicing, was low only for the acceptor site of exon II, which is preceded by a region of high information content that could contain sequences controlling splicing. This analysis explains previous observations of tissue-specific skipping during mRNA processing, resulting in the loss of the transcription start site in exon II, in human tissues.

Keywords: Pol H; Pol I; Xeroderma pigmentosum variant; Alternative splice; Donor; Acceptor; DNA damage

Several new classes of DNA polymerases have recently been identified in human cells, one being related to the bacterial class of mutagenic polymerases involved in the SOS repair system [1]. These class Y polymerases have reduced fidelity and are able to replicate a variety of damaged DNA templates with relaxed specificity [2–5]. The catalytic regions of the genes specifying these polymerases are in many cases homologs of the catalytic regions of the bacterial UMUC'D polymerase [1]. The catalytic regions have larger active sites than the replicative class B polymerases (Pol α, δ, ϵ) and can accommodate damaged bases or covalent adducts on template strands [6,7].

Recent work on these polymerases was stimulated by the discovery that the gene for the xeroderma pigmentosum variant (XP-V), a human disorder exhibiting high levels of UV-induced carcinogenesis, was a DNA polymerase [2,4]. The XP-V gene, *hRad30A*, Pol η or *POLH*, on chromosome 6p21, has a paralog in the human genome, *hRad30B* or *POLI*, on chromosome 18 [8–10], but only a single copy is found in yeast, *yRad30* [11]. The two genes represent ancient duplications that produce polymerases with overlapping specificities for replicating damaged DNA that can partially compensate for one another, as is commonly found for many genes in eukaryotic cells [12]. We previously reported that *POLH* undergoes significant amounts of alternative splicing; in the testis and fetal liver exon II, which encodes the ATG start site, is frequently spliced out [13]. We therefore conducted a detailed analysis of the genomic regions containing the *POLH* and *POLI* genes in human and mouse and determined the splicing efficiency in the regions of each intron/exon junction in human *POLH*. This approach can be used to understand the consequences of gene duplication to produce paralogs and the causes of exon skipping.

Genome scanning techniques such as comparative

[*] Corresponding author. Fax: +1-415-476-8218.
E-mail address: jcleaver@cc.ucsf.edu (J.E. Cleaver).

0888-7543/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0888-7543(03)00180-0

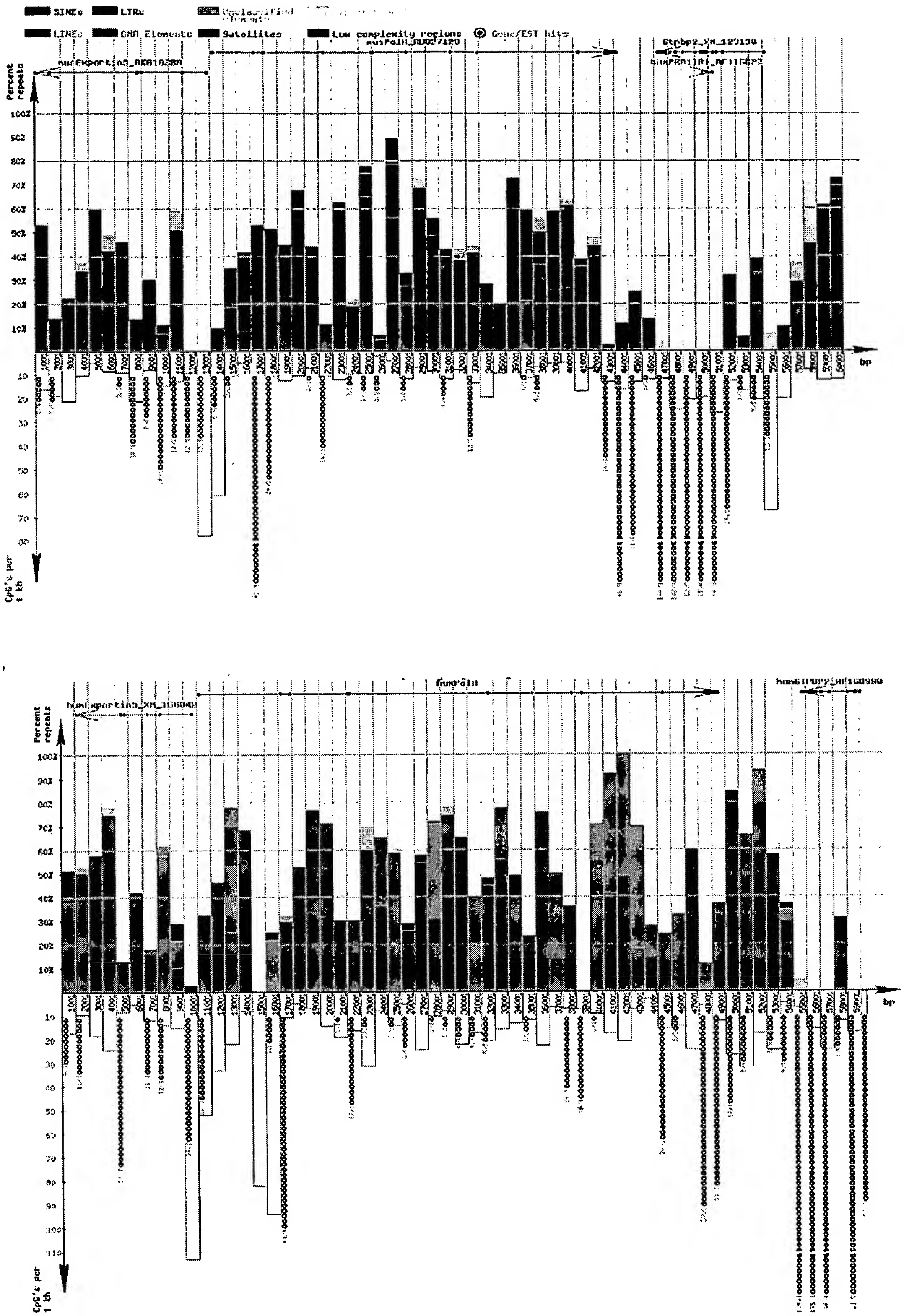


Fig. 1. Genome Cryptographer analysis of mouse (top) and human (bottom) *POLH*. Histograms of colored bars are shown at 1-kb intervals across the genomic regions containing *POLH*. The legend for the color of each bar is shown at the top. The gene is indicated above the histogram, annotated to show exons as small blocks on a line denoting the span of the gene. CpG regions are shown as histograms below the *x* axis line. Orange circles denote ESTs.

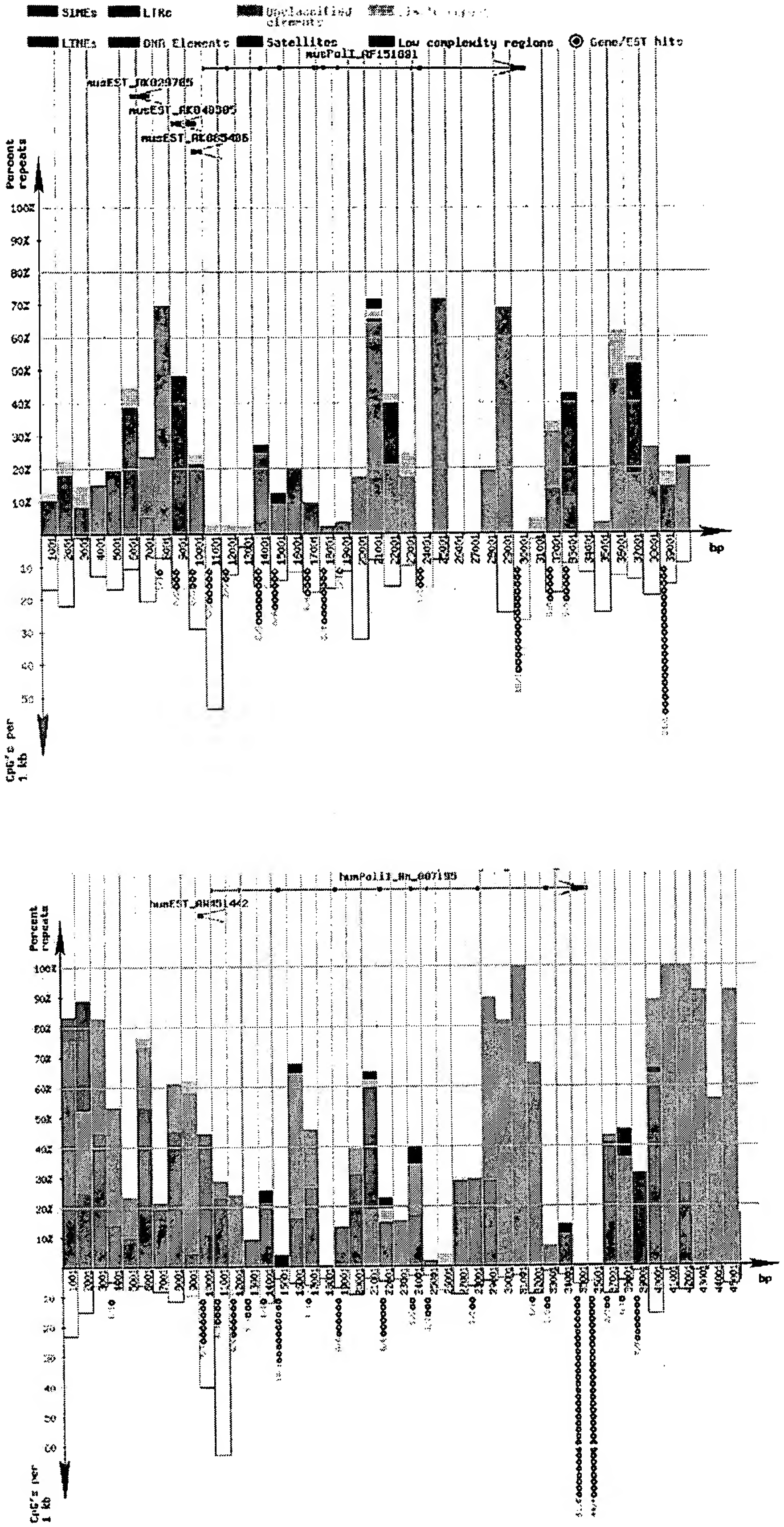


Fig. 2. Genome Cryptographer analysis of mouse (top) and human (bottom) *POLI*. Histograms of colored bars are shown at 1-kb intervals across the genomic region containing *POLI*. The legend for the color of each bar is shown at the top. The gene is indicated above the histogram, annotated to show exons as small blocks on a line denoting the span of the gene. CpG regions are shown as histograms below the *x* axis line. Orange circles denote ESTs.

genomic hybridization (CGH), restriction landmark genome scanning, and analysis of loss of heterozygosity have mapped numerous regions of recurrent genome copy number abnormalities in human solid tumors [14]. Although these mapping techniques have more often been used in analysis of genomic changes associated with malignancy, they can also be used for detailed analysis of individual genes in their native state. We used the suite of software tools collectively called Genome Cryptographer (GC) to facilitate integrative analysis. GC collects genome sequence information from multiple databases and visually displays it in analysis intervals (AIs) of constant width along the genome. Displayed information includes CpG density, sequence-tagged sites, expressed sequence tag (EST) clusters, locations and densities of repeated sequences (e.g., *Alus*, SINEs, LINEs), duplicons, similarities with syntenic murine sequences, known genes, and genome copy number determined using array CGH. This analysis produces a detailed map of the genomic landscape of these regions. We previously applied GC to the analysis of 1.2 Mb of 20q13.2 because it is amplified in a wide range of tumor types [14–16]. We have now applied it to the analysis of the chromosomal regions carrying *POLH* and *POLI* and also applied information theory to the analysis of the fine structure of intron/exon junctions. This has enabled us to map splice sites and their strengths and coordinate this information with the GC analysis to explain some of the details of *POLH* gene expression.

Results and discussion of GC sequence analyses

The results obtained from GC analysis for human and mouse *POLH* and *POLI* at 1-kb intervals across their respective chromosomal regions revealed the distribution of repetitive elements in intronic regions and regions of low complexity corresponding to many of the exons (Figs. 1 and 2). Coding exons appear as valleys in these distributions, repetitive regions are represented by colored bars. *POLH* appeared to contain more SINE elements (red bars) (Fig. 1) and *POLI* more satellite regions (green bars) (Fig. 2). CpG islands were evident at the 5′ end of both genes and additionally in the intron I region of *POLH*. Although *POLH* and *POLI* have a common origin before the evolution of the mammalian clade, based on their primary coding sequence, the intronic regions have undergone considerable change in their distribution of repetitive elements. Although each polymerase gene retains considerable similarity across species, the two polymerases have diverged from one another, as expected for paralogs resulting from gene duplication and evolutionary drift [12].

In our initial analysis a valley in intron IV (position 26000, Fig. 1, bottom) of the human *POLH* appeared to correspond the first exon of a gene, *exportin-5*, that was transcribed in the opposite direction [13,17]. *Exportin-5* is involved in nuclear export of double-stranded RNA binding proteins [17]. Subsequently the assembly of this region was changed. The *exportin-5* gene now appears not to overlap with *POLH*. In the human genome a space of about 2 kb lies between the 5′ ends of the *exportin-5* and *POLH* genes; in the mouse there is negligible space between the two genes. The valley in exon IV suggests that a residual or pseudoexon may remain in this region. The promoters of *POLH* and *exportin-5* are still likely to overlap in both human and mouse genomes. Deletion of the 5′ end of *PolH* in initial attempts to make a *PolH* knockout mouse resulted in embryo lethals possibly because of simultaneous deletion of regions of both *exportin-5* and *PolH*. Targeting the 3′ end of the gene should therefore be more successful in making a viable knockout mouse, because human patients who have chain-terminating mutations that result in no protein being synthesized are viable [18].

The architecture of the *POLH* genomic region is typical of many human genes, especially those associated with human disease, in having an untranslated first exon, a long first intron, and juxtaposed genes transcribed in opposite directions [19]. Of particular interest is the intron I region of human *POLH*, because we have previously observed tissue-specific splicing that eliminated exon II [13]. This region contains approximately 1 to 2 kb of very low complexity sequence and a high frequency of CpG sequences (Fig. 1, bottom). It is likely therefore that this region may have importance in regulating the efficiency of splicing and subsequent expression of an alternatively spliced variant of *POLH* lacking exon II. This exon contains the translation start site and therefore alternative splicing may be a mechanism of posttranscriptional regulation via inactivation of the message.

The murine *PolH* genomic region resembles human *POLH* in its pattern of repetitive elements (Fig. 1, top). There are regions of low complexity at the 5′ and 3′ ends of the gene and aligning with some but not all the coding exons. There is, however, no low-complexity region in intron I in the mouse gene. We therefore predict that elimination of exon II by alternative splicing that has been observed in specific human tissues such as lung, testis, and embryonic liver will be less likely to occur in mouse tissues [13].

Results and discussion of splice site analyses

Splicing efficiency is determined by the sequence context around each individual splice site. To analyze splicing efficiency in human *POLH*, to understand the regulation of the gene in the region of alternative splicing, we have calculated the "individual information variable" (R_i, Methods Eq. (2)). This value represents the probability that a site acts as an acceptor or donor in splicing. In theory, R_i values

Table 1
Splice site R_i values and locations and exon locations through the *POLH* gene

Type	Value	Location (Donor)	Location (Acceptor)	Type	Value	Location (Donor)	Location (Acceptor)
Donor	5.7 bits	13965		**< Exon**	**No. 7**	**41431**	**41550**
> Exon	**No. 1**	**13747**	**13979**	Acceptor	5.3 bits	41436	
> Donor	7.4 bits		13980	Donor	3.0 bits		41493
Acceptor	3.8 bits	14018		Acceptor	3.0 bits	41521	
Acceptor	2.9 bits	14057		Acceptor	4.5 bits	41527	
Donor	2.5 bits		14060	Acceptor	3.0 bits	41521	
Acceptor	3.5 bits	19817		Acceptor	4.5 bits	41527	
Donor	3.2 bits		19818	Acceptor	2.9 bits	42107	
< Acceptor	3.4 bits	19854		Acceptor	3.3 bits	42143	
< Exon	**No. 2**	**19855**	**19995**	**> Exon**	**No. 7**	**41431**	**41550**
Acceptor	4.4 bits	19919		> Donor	9.2 bits		41551
Donor	2.6 bits		19989	Donor	3.5 bits		41555
Donor	6.5 bits		19992	< Acceptor	7.6 bits	42153	
> Exon	**No. 2**	**19855**	**19995**	**< Exon**	**No. 8**	**42154**	**42277**
> Donor	5.7 bits		19996	Donor	4.8 bits		42234
Donor	4.7 bits		20031	**> Exon**	**No. 8**	**42154**	**42277**
Acceptor	4.4 bits	20540		> Donor	10.8 bits		42278
< Acceptor	11.3 bits	20545		Acceptor	4.4 bits	42317	
< Exon	**No. 3**	**20546**	**20680**	Donor	4.9 bits		42320
Donor	5.6 bits		20556	Donor	3.1 bits		42324
Donor	5.8 bits		20593	Donor	5.7 bits		42328
Donor	2.5 bits		20646	Donor	2.9 bits		42338
Donor	4.4 bits		20650	Acceptor	3.2 bits	42746	
> Exon	**No. 3**	**20546**	**20680**	< Acceptor	4.7 bits	42792	
> Donor	8.1 bits		20681	**< Exon**	**No. 9**	**42793**	**42858**
Donor	2.7 bits		20750	Donor	2.4 bits		42793
Donor	5.9 bits		20770	Donor	3.9 bits	42801	
Donor	6.1 bits		24739	**> Exon**	**No. 9**	**42793**	**42858**
< Acceptor	5.7 bits	24810		> Donor	6.1 bits		42859
< Exon	**No. 4**	**24811**	**25028**	Acceptor	6.5 bits	42883	
Donor	3.3 bits		24811	Acceptor	3.8 bits	42884	
Donor	3.4 bits		24833	Donor	4.4 bits		42886
> Exon	**No. 4**	**24811**	**25028**	Acceptor	4.4 bits	42933	
> Donor	5.4 bits		25029	Acceptor	3.0 bits	48086	
Acceptor	3.3 bits	25069		Acceptor	3.5 bits	48090	
Donor	8.3 bits		35213	< Acceptor	11.4 bits	48092	
< Acceptor	5.5 bits	35234		**< Exon**	**No. 10**	**48093**	**48262**
< Exon	**No. 5**	**35235**	**35404**	Donor	4.3 bits		48120
Acceptor	8.8 bits	35281		**> Exon**	**No. 10**	**48093**	**48262**
Acceptor	7.3 bits	35303		> Donor	9.5 bits		48263
Acceptor	7.4 bits	35311		Acceptor	3.1 bits	48278	
> Exon	**No. 5**	**35235**	**35404**	Acceptor	10.4 bits	48291	
> Donor	7.8 bits		35405	Acceptor	5.4 bits	48296	
Donor	3.7 bits		35460	Acceptor	4.0 bits	48321	
Donor	2.7 bits		35469	Acceptor	4.8 bits	51193	
< Acceptor	9.3 bits	38526		Acceptor	3.8 bits	51196	
< Exon	**No. 6**	**38527**	**38630**	< Acceptor	12.1 bits	51198	
Donor	3.8 bits		38589	**< Exon**	**No. 11**	**51199**	**53181**
Acceptor	2.8 bits	38601		Acceptor	3.4 bits	51213	
Acceptor	2.7 bits	38609		Acceptor	3.8 bits	51236	
Acceptor	6.3 bits	38677		Acceptor	2.9 bits	51260	
> Exon	**No. 6**	**38527**	**38630**	Acceptor	4.3 bits	51267	
> Donor	9.3 bits		38631	Acceptor	12.5 bits	51269	
Acceptor	3.5 bits	41384		Acceptor	2.9 bits	51279	
< Acceptor	9.6 bits	41430		**Exon**	**No. 11**	**51199**	**53181**

Actual exon splice site positions are indicated by boldface; individual donor and acceptor sites marked by > and <, respectively.

of at least zero are required for an acceptor or donor site to exist. Empirically, it has been found that an R_i value of at least 2.4 bits is almost always required for a splice site to be functional [20]. Strong acceptor sites are in the range of 9 to 10 bits and up. Strong donor sites are in the range of 7 or 8 bits and up.

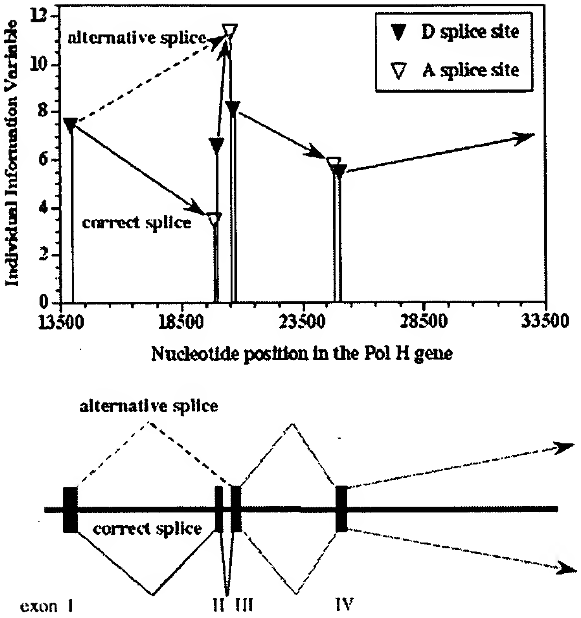


Fig. 3. Top: Splice site R_i values shown for empirically determined splice sites in the first four exons of human *POLH*. Solid inverted triangles, donor sites; open inverted triangles, acceptor sites; arrows show direction of correct (solid arrows) and alternative (dashed arrow) splicing between donor and acceptor sites between exon I and exon II. Bottom: First four exons of *POLH* gene with location of exons shown by solid rectangles and splicing shown by connecting lines between donor and acceptor sites. Alternative splicing causes skipping of exon II.

Wild-type POLH

We analyzed the complete genomic region of human *POLH* for the occurrence of all possible donor and acceptor sites and calculated their R_i values. Many values were in the range of 2.5 to 5.5, which is below that usually required for effective splicing (Table 1). Most of the splice sites that correspond to empirically known splice sites had R_i values that indicated strong splicing. These values ranged from 4.5 to 12.1 bits and neighboring sites generally had lower values (Table 1, Fig. 3). The major exception was the acceptor site for exon II that had a low value of 3.5 bits and had neighboring sites with comparable to larger values (Table 1, Fig. 3). This site is skipped in a number of tissues and to some extent in cell culture, in preference for the acceptor site of the next exon, III, resulting in the loss of 141 nt of exon II [13].

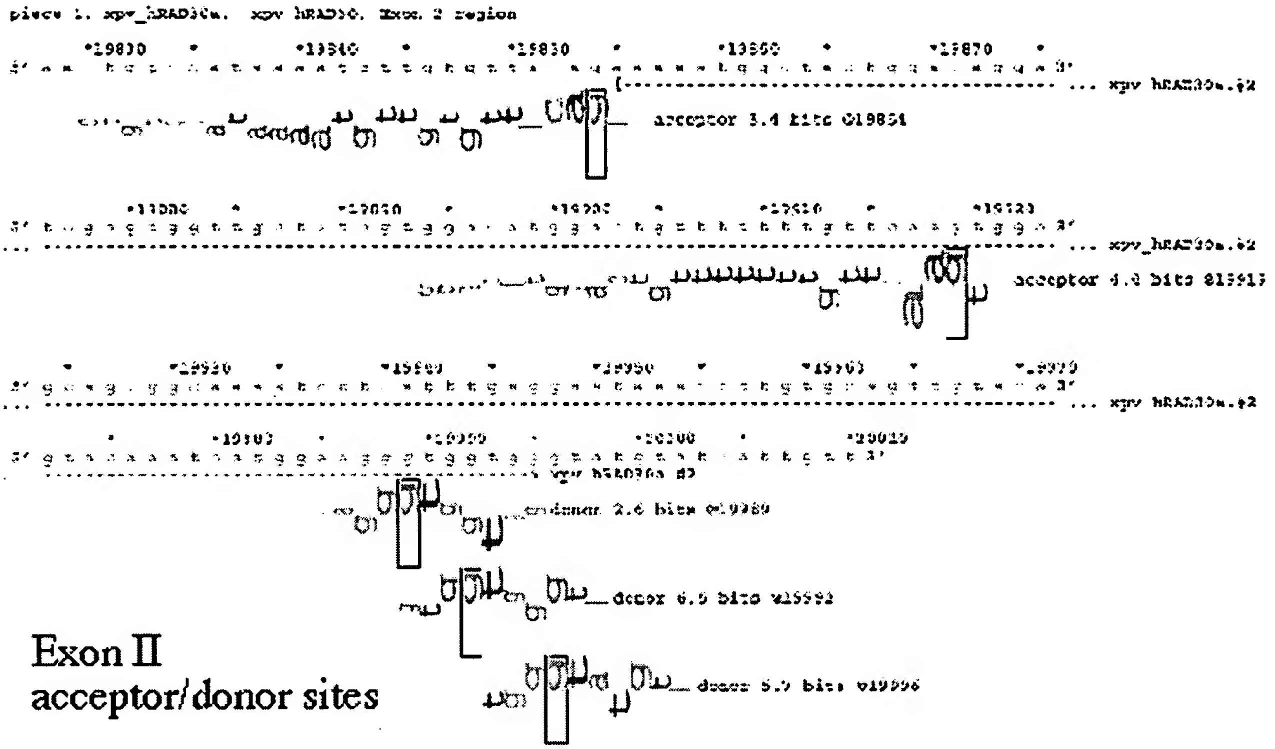
The acceptor site of strength 3.5 bits at location 19854 is weak but in the correct location to introduce exon II beginning at 19855 (Fig. 4). There is an acceptor with R_i = 3.5 bits at 19817 and one with R_i = 4.4 bits at 19919. These are all relatively weak acceptors, but above the absolute limit of 0 bits for definition of a site and above the empirical limit of 2.4 bits for functional sites. There is a donor site with R_i = 5.7 bits at 19996, corresponding to the end of exon II at 19995. However, there are also donors with R_i = 6.5 bits at 19992 and 4.7 bits at 20031. These are moderately strong sites. There is even a fairly weak one at 19989 with R_i = 2.6 bits. These sites, along with observed exon II, are illustrated in Fig. 4.

A good starting point for relating individual information R_i values to binding effectiveness is to regard the binding site affinity, and thus the likelihood of use, of a site with value R_i as proportional to $2\hat{}(R_i)$. With acceptors this weak, there could be a significant fraction of exon skipping, as has been observed experimentally [13]. There could be as many as a half-dozen alternative exons in this situation, with binding site affinities within a factor of 2 to 4 of each other. This grows to nine sites with affinities within another factor of 2, if the site at 20031 is used. If the donor at 19989 is active, up to a dozen alternatives are possible. These possible alternative exons are shown with the splice sites in Fig. 4.

Effect of mutations in POLH on splicing efficiencies

Several XP-V cases have now been analyzed and mutations have been identified in the N-terminal catalytic domain and at the C-terminal end, which regulates nuclear foci formation and PCNA interaction [2,4,21–23]. Yuasa et al. [21] report that a G → C mutation at location 19854, which is at the end of intron I, just preceding exon II, results in skipping exon II. Our analysis finds that this mutation changes the R_i value of this acceptor from 3.5 to −3.8 bits (Table 2). Thus, by our analysis, this mutation changes a weak, but adequate, site to one that is almost certainly not functional. This is consistent with skipping exon II, as reported. Other mutations in various XP-V patients occur sufficiently far from the nearest splice sites that they made

Fig. 4. Top: Walker analysis of the wild-type genomic sequence of human *POLH* in the area of exon II. Genomic sequences are shown horizontally, with locations given above each in increments of 10 bp. Asterisks indicate locations that are multiples of 5. A brief description of each piece of DNA is given above the locations. Individual information contributions are shown below the sequence, with positive contributions pointing up and negative contributions pointing down. The positions of splice sites are boxed. The sites are labeled with type, strength (R_i value), and location. Exon II is shown as a horizontal dashed line between "[" and ">" symbols. It is initiated by an acceptor site and terminated by a donor site. Bottom: Walker analysis of the wild-type genomic sequence of human *POLH* in the area of exon II, with predictions of possible alternative splice exons. Possible alternative exons consistent with the splice sites found are shown as horizontal dashed lines between "[" and ">" symbols. They are initiated by acceptor sites and terminated by donor sites. For this set, the acceptors have strengths of at least 3.4 bits, and the donors have strengths of at least 4.7 bits. Here, exon II specified in the data set is predictable as 2(b2), and eight other putative alternative exons are indicated. These are coded by piece, exon, acceptor beginning, and donor ending. For example, alternative exon 2(b3) begins after acceptor "b", at location 19855, and ends before donor "3", at location 20030. In this case, "b" and "3" are simply ordering conventions.


piece 1. xpv_hRAD30a. xpv hRAD30. Exon 2 region
*19830
*19840
*19850
*19860
*19870
xpv hRAD30a.#2
acceptor 3.4 bits @19851
acceptor 4.8 bits @19917
xpv hRAD30a #2
donor 2.6 bits @19989
donor 6.5 bits @19992
donor 5.9 bits @19996
*20010
Exon II
acceptor/donor sites

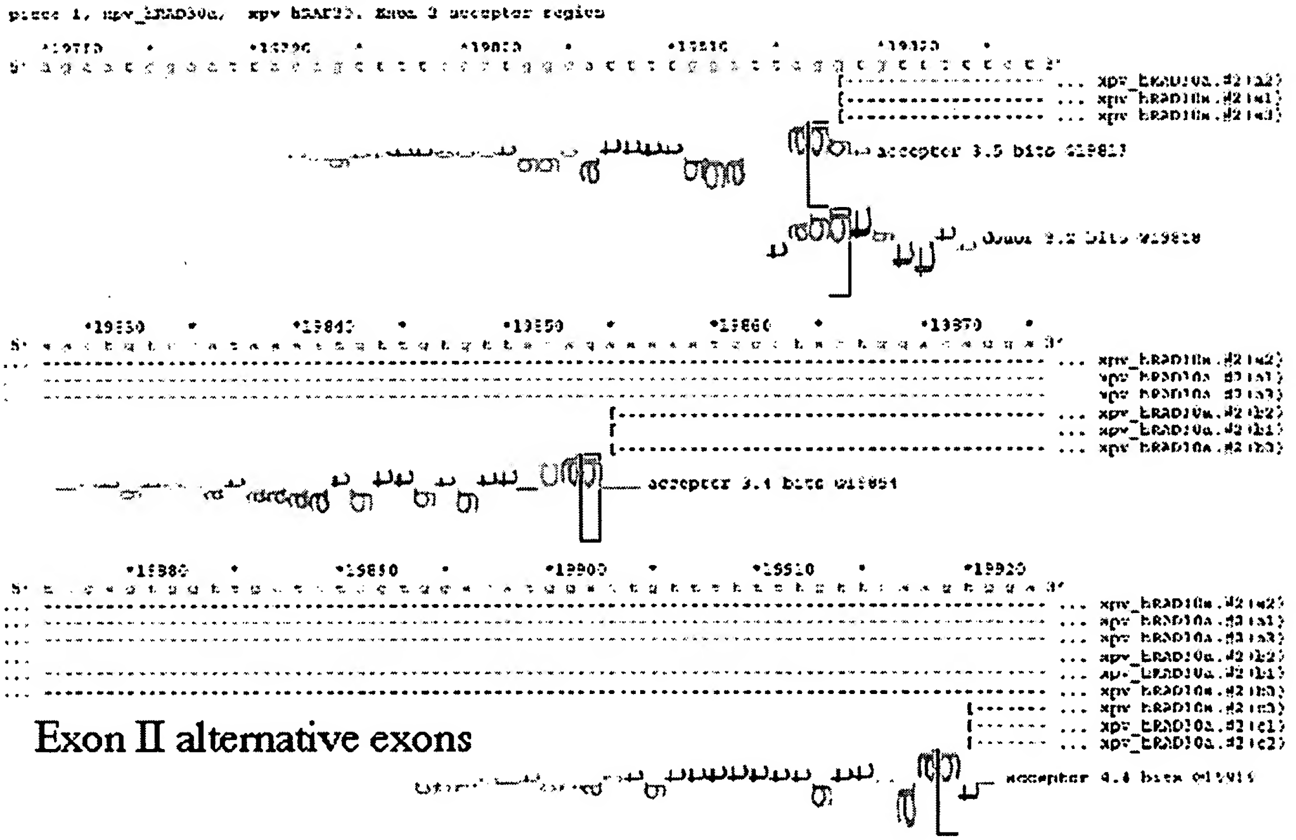

piece 1, xpv_hRAD30a, xpv hRAD30. Exon 2 acceptor region
acceptor 3.5 bits @19813
donor 2.2 bits @19818
*19830
*19840
*19850
*19860
*19870
acceptor 3.4 bits @19851
*19880
*19890
*19900
*19910
*19920
acceptor 4.4 bits @19915
Exon II alternative exons

Table 2
Effect of mutations on splice site R_i values in adjacent potential splice sites

Cell line	Exon	Mutation	Splice site	R_i change[a]	Effect
XP1RO	2	g19854a	19854	A 3.4 to −3.8	Splice site lost
	8	g42185t	42153	A 7.6 to 7.6	Still used
			42160	None to A 3.2	Not used
XP30RO	2	19964 del 13	19989	D 2.6 to 2.6	Not used
			19992	D 6.5 to 6.5	Still used
XP7TA	5	35276 del 2	35281	A 8.8 to 7.9	Not used
			35303	A 7.3 to 7.3	Not used
XP2SA	8	g42185t	42207	A 0.5 to 1.4	Not used

[a] A represents an acceptor site, D a donor site.

small but insignificant changes in the acceptor or donor splicing efficiencies (Table 2).

Final comments

This analysis by two independent approaches provides a sequence-based interpretation of the observed alternative splicing seen in human *POLH* and comparisons between the two paralogs *POLH* and *POLI* in human and mouse. The analysis suggests that further experiments could profitably search for regulatory elements and binding proteins within the intron I region of low complexity. The analysis further indicates that murine *PolH* does not have the same low-complexity region in intron I as seen in human *POLH*. If this region plays a significant role in regulation of alternative splicing, then exon II should be skipped only in human and not mouse. Experiments to test this are under way. This intronic region, coupled with the low R_i value for the acceptor site of exon II explains the observed loss of exon II in a significant number of transcripts. In the testis the alternative splicing appears particularly high, involving loss of exon II in almost half the transcripts [13]. This may represent a mechanism that partially down regulates the low-fidelity polymerase to permit increased activity of recombination pathways that we have found to be up regulated in the absence of Pol H [24].

Methods

Genome Cryptographer analysis.

GC is a suite of Perl programs designed to facilitate megabase scale analysis of genomic sequence [14]. This suite is built of separate modules that exchange information via intermediate text files. Data in intermediate files are written in a consistent format: sequence name, sequence length, window size, appropriate data for a given window (the number of these "data" lines equals the number of windows that are contained per sequence), and, optionally, after a blank line, annotation data.

Analysis of the sequence is done in the following stages: Using script gc_plot.pl, we generate the plot of the GC content and number of CpG dinucleotides per AI. The CpG dinucleotide density is weighted by adding 0.25 to the dinucleotide count for each CpG dinucleotide that is found within 20 bp of another. This makes CpG islands more apparent as peaks in CpG dinucleotide density plots. The script also produces the graphic plot of the GC and CpG content and, if available, can annotate the plot with features from the output of the count_gene.pl script (making it easier to correlate changes in GC and CpG content with sequence features).

The sequence is analyzed for repeats using the publicly available RepeatMasker program (Smit and Green, http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl). RepeatMasker output files are saved. Masked sequence is used for searches of public and proprietary databases. Currently, GC employs the NCBI version of BLAST (ftp://ncbi.nlm.nih.gov/blast/). Sequence is compared to nonredundant, HTGS, dbSTS, and dbEST divisions of GenBank. Sequence similarity criteria are set to reduce the probability of identifying ESTs from members of closely related gene families (cutoff of expect score 10–20).

Optionally, masked sequence is searched against a database containing syntenic sequences of model organisms (in our case, *PolH* mouse sequence from syntenic region of mouse chromosome 17 [13]). Count_gene.pl and count_homol.pl are used to analyze output of the BLAST searches, creating a list of the number of relevant hits per AI. Count_gene.pl also generates a first draft of sequence annotation data, by capturing all the database hits that exceed a user-selectable threshold in length. If desired, this annotation can be extended and updated by the user manually. We captured the exact coordinates of regions of identity of database hits used for annotation. This information proved to be invaluable for analysis of the gene relationships, because the alignment of cDNA sequence to genomic sequence automatically yields intron–exon organization of the corresponding gene.

Finally, graph.pl is used to gather information produced by gc_plot.pl (CpG distribution data), RepeatMasker (repeat distribution data), count_gene.pl (annotation and distribution of database hits), and count_homol.pl (distribution of conserved regions) and produce a graphical summary. Currently we are working on the extension of graph.pl capabilities (to make output interactive and to add capability to include gene expression and copy number data from array-based experiments). The first version of the Genome Cryptographer software is accessible at http://kinase.ucsf.edu/gc.

Splice site analysis

Computer analysis of the strength of splice sites was performed using programs from the Individual Information package of T. Schneider [25]. Values of the individual information variable $R_i(b, l, j)$ are calculated for each base and position of the selected sequence, j, in the domain of interest. In this situation, $R_i(b, l, j)$ is the difference in uncertainty before and after binding of a specific base, b, at a specific position, l, relative to the origin of an acceptor or donor sequence, j.

These uncertainties are based on probability estimates. Probabilities are estimated from the relative frequencies, $f(b, l)$, of occurrence of bases, b, at specified positions, l, relative to the splice site origin in a set of known splice sites. Weighting matrices based on the relative frequencies of bases within a specified domain of the location have been constructed from a collection of more than 1700 aligned sequences from known acceptor and donor sites [26]. Entries of the weighting matrices are

$$R_{iw}(b, l) = 2 - (-\log_2(f(b, l))) + e(n, l), \qquad (1)$$

where $e(n, l)$ is a small-sample error correction for n samples at position l.

The R_i value of a site at a selected location, j, is the sum of the individual R_i values of a sequence of bases over a restricted domain about that location. Symbolically,

$$R_i(j) = \Sigma_{l=\mathrm{site\_domain}} \Sigma_{b=\{\mathrm{acgt}\}} s(b, l, j)\, R_{iw}(b, l), \qquad (2)$$

where the function $s(b, l, j) = 1$ specifies base b at position l for sequence j and is 0 otherwise. For acceptors in human DNA the site domain is -25 to $+2$; for donors, it is -3 to $+6$. Given any specific genomic sequence, the site sequence is determined entirely by the position of an origin, or other offset, of the site in the genomic sequence.

The values of R_i are normally expressed in bits (binary digits). One bit is the amount of information needed to distinguish between two choices, 2 bits are needed to choose one of four choices, etc. Reasons for choosing this functional form and these units are discussed in Schneider [25] and Shannon [27,28].

Sequence walkers illustrate graphically each R_i contribution of the bases to the acceptor or donor site at a location [29]. On these plots, bases that contribute positively to the R_i sum point upward, those that contribute negatively point down. The height of the letter is proportional to the information contribution of that base. The R_i sum is given along with the type of site for which it is calculated.

For more information on individual information, sequence logos, information theory, and related topics, see the Schneider Lab Web page, http://www.lecb.ncifcrf.gov/~toms/.

Programs used were mkdb, dbbk, catal, delila, scan, exon, and lister. These are a few of the members of the large DELILA package [29]. The programs were initially run on a Sun SPARCstation 2 under Sun OS 5.5. Recent runs were done on a Sun Blade 1000 using Solaris 8.

Mkdb, dbbk, and catal convert genomic DNA data from GenBank format to delila format. Delila processes the data, selecting pieces and providing mutations. Scan locates splice sites with specified properties. Lister prepares the walker plots with splice sites and exons displayed. The panels of Fig. 4 were generated this way.

Fig. 3 gives a different perspective. For clarity of display, the bit numbers were displayed as vertical bars at the appropriate sites through the genomic regions, and only those values occurring at the sites of known splice sites were used.

Web site references

Genome Cryptographer: http://kinase.ucsf.edu/gc. Collins and Volk.

NIH BLAST Web site: ftp://ncbi.nlm.nih.gov/blast/.

RepeatMasker Web page: http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl. Smit and Green.

Schneider Laboratory Web page: http://www.lecb.ncifcrf.gov/~toms/. Molecular information theory and the theory of molecular machines.

Acknowledgments

The work described here was supported by National Institutes of Environmental Health Sciences Grant 1 RO1 ES 8061 (to J.E.C.) and the Ellison Senior Scholar Award (to J.E.C.) and the UCSF Breast and Prostate SPORE programs (to C.C. and S.V.). We are also grateful to the XP Society (Poughkeepsie, NY), for their continued support and encouragement.

References

[1] H. Ohmori, et al., The Y-family of DNA polymerases, Mol. Cell 8 (2001) 7–8.

[2] R.E. Johnson, C.M. Kondratick, S. Prakash, L. Prakash, *hRAD30* mutations in the variant form of xeroderma pigmentosum, Science 264 (1999) 263–265.

[3] R.E. Johnson, M.T. Washington, S. Prakash, L. Prakash, Fidelity of human DNA polymerase η, J. Biol. Chem. 275 (2000) 7447–7450.

[4] C. Masutani, et al., The *XPV* (xeroderma pigmentosum variant) gene encodes human DNA polymerase η, Nature 399 (1999) 700–704 DOI: 10.1038/21447..

[5] T. Matsuda, K. Bebenek, C. Masutani, F. Hanoaka, T.A. Kunkel, Low fidelity DNA synthesis by human DNA polymerase eta, Nature 404 (2000) 1011–1013, doi:10.1038/35010014.

[6] J. Trincao, et al., Structure of the catalytic core of S. cerevisiae DNA polymerase η: Implications for translesion synthesis, Mol. Cell 8 (2001) 417–426.

[7] H. Ling, F. Boudsocq, R. Woodgate, W. Yang, Crystal structure of a Y-family DNA polymerase in action. A mechanism for error-prone and lesion-bypass replication, Cell. 107 (2001) 91–102.

[8] A. Tissier, J.P. McDonald, E.G. Frank, R. Woodgate, Pol iota, a remarkable error-prone human DNA polymerase, Genes Dev. 14 (2000) 1642–1650.

[9] A. Tissier, et al., Misinsertion and bypass of thymine–thymine dimers by human DNA polymerase iota, EMBO J. 19 (2000) 5259–5266.

[10] A. Tissier, et al., Biochemical characterization of human DNA polymerase iota provides clues to its biological function, Biochem. Soc. Trans. 29 (2001) 183–187.

[11] R.E. Johnson, S. Prakash, L. Prakash, Efficient bypass of a thymine–thymine dimer by yeast DNA polymerase eta, Science 283 (1999) 1001–1004.

[12] Z. Gu, et al., Role of duplicate genes in genetic robustness against null mutations, Nature 421 (2002) 63–66, doi:10.1038/nature01198.

[13] M. Thakur, et al., DNA polymerase H undergoes alternative splicing, protects against UV sensitivity and apoptosis, and suppresses Mre11-dependent recombination, Genes Chromosomes Cancer 32 (2001) 222–235.

[14] C. Collins, et al., Comprehensive genome sequence analysis of a breast cancer amplicon, Genome Res. 11 (2001) 1034–1042, doi: 10.1101/gr.174301.

[15] A. Kallioniemi, et al., Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors, Science 258 (1992) 818–821.

[16] C. Collins, et al., Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma, Proc. Natl. Acad. Sci. USA 95 (1998) 8703–8708.

[17] A.M. Brownawell, I.G. Macara, Exportin-5, a novel karyopherin, mediates nuclear export of double-stranded RNA binding proteins, J. Cell Biol. 156 (2002) 53–64, doi: 10.1083/jcb.200110082.

[18] B.C. Broughton, et al., Molecular analysis of mutations in DNA polymerase eta in xeroderma pigmentosum-variant patients, Proc. Natl. Acad. Sci. USA 99 (2002) 815–820 doi: 10.1073/pnas.022473899.

[19] S. Karlin, C. Chen, A.J. Gentles, M. Cleary, Associations between human disease genes and overlapping gene groups and multiple amino acid runs, Proc. Natl. Acad. Sci. USA 99 (2002) 17008–17013, doi: 10.1073/pnas.262658799.

[20] P.K. Rogan, B.M. Faux, T.D. Schneider, Information analysis of human splice site mutations, Hum. Mutat. 12 (1998) 153–171.

[21] M. Yuasa, C. Masutani, T. Eki, F. Hanaoka, Genomic structure, chromosomal localization and identification of mutations in the xeroderma pigmentosum variant (XPV) gene, Oncogene. 19 (2000) 4721–4728, doi: 10.1038/sj.onc.1203842.

[22] P. Kannouche, et al., Domain structure, localization, and function of DNA polymerase eta, defective in xeroderma pigmentosum variant cells, Genes Dev. 15 (2001) 158–172.

[23] L. Haracska, C.M. Kondratick, I. Unk, S. Prakash, L. Prakash, Interaction with PCNA is essential for yeast DNA polymerase η function, Mol. Cell 8 (2001) 407–415.

[24] C.L. Limoli, E. Giedzinski, W.F. Morgan, J.E. Cleaver, Polymerase η deficiency in the XP variant uncovers an overlap between the S phase checkpoint and double strand break repair, Proc. Natl. Acad. Sci. USA 97 (2000) 7939–7946, doi: 10.1073/pnas.130182897.

[25] T.D. Schneider, Information content of individual genetic sequences, J. Theor. Biol. 189 (1997) 427–441.

[26] R.M. Stephens, T.D. Schneider, Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites, J. Mol. Biol. 228 (1992) 124–1136.

[27] C.E. Shannon, A mathematical theory of communication (part I), Bell Syst. Tech. J. 27 (1948) 379–423.

[28] C.E. Shannon, A mathematical theory of communication (part II), Bell Syst. Tech. J. 27 (1948) 623–656.

[29] T.D. Schneider, Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences, Nucleic Acids Res. 25 (1997) 4408–4415.